DOCUMENT RESUME

ED 189 530 CG 014 555

AUTHOR Strupp, Hans H.: And Others
TITLE Toward the Measurement of Negative Effects in Psychotherapy.
PUB DATE 19 Jun 80
NOTE 13p.: Paper presented at the Conference of the Society for Psychotherapy Research (Asilomar, CA, June 19, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Counseling Effectiveness: *Counseling Techniques: *Counselor Client Relationship: Item Analysis: *Measures (Individuals): Patients: Program Evaluation: *Psychotherapy: Rating Scales: Test Construction: *Test Reliability
IDENTIFIERS *Vanderbilt Negative Indicators Scale

ABSTRACT

The Vanderbilt Negative Indicators Scale (VNIS), an instrument designed to quantify characteristics of the patient, the therapist, and their interactions, was developed as part of an investigation of negative effects in psychotherapy by the Vanderbilt Psychotherapy Research Team. Revision of the original 25-item scale has proceeded in two major stages, i.e., an intensive review of project files to create additional VNIS items and the development of items to preserve the clinical relevance and inter-rater reliability of the instrument. Research team members (N=3) and two clinically-sophisticated judges rated therapy sessions from the project files to assess the inter-rater reliability of the revised 42-item VNIS. Results suggest that the VNIS is a promising method for detecting signs of ineffective and harmful psychotherapy early in a therapeutic relationship. (Author/HLM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY Hans H. Strupp
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED189530

CG 014555

TOWARD THE MEASUREMENT OF NEGATIVE EFFECTS IN PSYCHOTHERAPY

Vanderbilt Psychotherapy Research Team*

This paper is a progress report of research aimed at the systematic investigation of negative effects in psychotherapy. Our work has been focused on the development of a process scale to detect signs of either ineffective psychotherapy (i.e., no change) or harmful psychotherapy (i.e., detrimental to a patient's psychological well-being and/or confidence in psychotherapy as a treatment modality). This research effort was initiated by the Vanderbilt Psychotherapy Research Team following a survey of eminent theorists, clinicians, and researchers and a review of the empirical literature on the problem of negative effects in psychotherapy (Strupp, Hadley, & Gomes-Schwartz, 1977).

On the basis of this preliminary work, Gomes-Schwartz (1978), in collaboration with members of the Vanderbilt Psychotherapy Research Team, began developmental work on an instrument designed to quantify characteristics of the patient, the therapist, and their interaction that might lead to negative outcomes. The first version of this process measure which has since come to be known as the "Vanderbilt Negative Indicators Scale" (VNIS) was composed of 25 items derived from factors identified as causes of negative outcomes by respondents in the earlier survey.

Paper presented at the Conference of the Society for Psychotherapy Research: Asilomar, California, June 19, 1980.

* The Vanderbilt Psychotherapy Research Team includes Hans H. Strupp, Lois Keithly, Karla Moras, Stephanie Samples, Janet Sandell, and Gloria Waterhouse. Since the development of this instrument represents a group effort shared by all team members, collaborators are listed in alphabetical order.

Several scales of the original instrument were shown to differentiate between one patient who had improved and one who had deteriorated in the Vanderbilt Project. However, it was apparent that a great deal of further work was necessary before the VNIS would be ready for rigorous empirical application.

Revision has proceeded in two major stages. First, therapy sessions from the files of the Vanderbilt Project were extensively reviewed in order to detect patient, therapist, and interaction variables that distinguished good and poor outcome cases. These observations were translated into additional VNIS items. Secondly, considerable effort has been devoted to creating items that both preserve the clinical relevance of the instrument and that can be reliably rated by judges working independently. Practically, these two goals call for item definitions that tie commonly used clinical concepts to specific observables occurring in therapeutic interactions. Developing such definitions has been difficult because most clinical concepts, particularly psychodynamic concepts, do not have specific behavioral referents; rather, they are manifested by many different types or patterns of behavior. In addition, clinical concepts involve the attribution of meaning to behaviors, thus requiring inferences by raters. Consequently, recognition rules cannot be totally explicit and adequate inter-observer consensus is likely to be more difficult to achieve (cf. Duncan & Fiske, 1977).

The current version of the VNIS consists of 42 items. These are grouped into five sections, most of which reflect prevailing conceptualizations of the major outcome-related dimensions of the therapeutic interaction. The sections and the VNIS item categories are shown in Table 1.

Table 1

The VNIS rating instructions require two types of judgments. First, a judgment of presence or absence is made. A rating of "presence" is contingent upon a decision that a classifiable event or quality exceeds "normal limits": That is, it is evident to a greater degree than would be expected in an "average" therapy session or in the broad range of patients who are ordinarily offered out-patient psychotherapy. Any item judged present according to this criterion requires a second judgment, involving a numerical rating of frequency or intensity on a 1-5 scale.

Judges are presented with entire hours, rather than randomly selected segments of therapy sessions. We have come to believe that an entire session provides the requisite context for clinically-meaningful rating decisions. Either audio or video recordings of sessions provide an acceptable medium of presentation, although video recordings are considered preferable (see Kiesler, 1979, on process ratings of "relationship" variables).

For rating purposes, therapy sessions are subdivided into 10-minute segments and ratings are made following each segment. This procedure helps to keep the rater's attention focused on the task, and tends to reduce spurious variance introduced by memory and perceptual factors. Each item is given a final rating at the end of the session. These final ratings have been used in the preliminary analyses of inter-rater reliability and validity reported here.

Revised VNIS: Inter-rater Reliability and Validity

In an initial attempt to assess the inter-rater reliability of the revised VNIS, three members of the research team rated three therapy sessions, consisting of one session from each of three cases unfamiliar to team members. We were encouraged by the actual degree of consensus obtained for these three tapes, although we also observed that the usual statistical methods for computing inter-rater reliability, such as the Pearson correlation coefficient, seemed to underestimate inter-judge agreement. This was particularly true for sessions that judges viewed as basically "good," i.e., essentially lacking negative indicators, which resulted in low variability in ratings and consequently low r's.

In a subsequent field trial, two clinically-sophisticated judges received approximately 35 hours of training and then independently rated videotapes of the third therapy session for ten cases from the Vanderbilt Project. These cases consisted of one high change and one low change patient for each of five therapists. The therapists represented a range of theoretical orientations, including analytic and more eclectic approaches combining client-centered and Gestalt techniques.

Pearson r's were used to assess inter-rater reliability for each of the ten sessions rated in the field study. Correlations computed on all 42 items for each session ranged from .13 to .55. The correlation for ratings across all ten sessions (N=420) was .40. These reliabilities were lower than those obtained for four tapes independently rated during the training procedure, which ranged from .49 to .66. Comparison of the actual ratings with the obtained r's once again revealed that low

variance in ratings often resulted in low correlation coefficients that did not reflect the actual degree of consensus between raters. To explore the effect of other contributing factors, Finn's r statistic (Finn, 1970, 1972; Tinsley & Weiss, 1975) was used to assess inter-rater reliabilities for individual items. Reliabilities were not computed for eight items which were never rated. Rater agreement exceeded chance on 18 of the remaining 34 items. Finn's r's on these items ranged from .76 to .98 with a median r of .90.

Simple percent agreements calculated on rated items indicated that the judges often chose different items within a VNIS subsection to capture their perceptions of negative indicators. This led us to conclude that the most meaningful index of inter-judge agreement would be based on sums of reliably rated items within each VNIS subsection. These inter-rater reliabilities are presented in Table 2. The figures shown are Pearson correlations computed on the sum of each judge's item ratings in VNIS subsections for each session.

Table 2

High rater agreement was achieved on Patient Qualities and Therapist Qualities subsections. Similarly, there was good agreement between raters on total VNIS ratings. In contrast, the correlation of -.25 for Therapist Errors in Technique items suggests that subjective factors heavily influenced judgments of technique errors.

To examine test-retest reliability, intra-rater reliabilities were computed for three sessions rated twice by the same judge on different occasions. Pearson r's were moderate (.46) to high (.97), with a median correlation of .80.

Next, we assessed the power of VNIS total scores (exclusive of unreliable items) to discriminate between high and low outcome cases. t (4) = -2.87* was significant at the $p < .025$ level (one-tailed test). Since this analysis was performed on ratings for the third session of therapy only, the power of the VNIS to discriminate between high and low outcome cases seems particularly promising. Similar analyses were performed for total scores on each subsection of the VNIS. The difference between high and low outcome cases on the Patient Qualities section approached significance at the $p < .08$ level (t (4) = -1.69, one-tailed test). Differences between the outcome groups on the remaining subsection scores were not significant, although all means were in the right direction. Two observations are relevant to these analyses. First, the statistical tests were performed on very small samples. Secondly, there were large standard deviations that in some instances approached the size of the mean. Both factors make it difficult to attain the traditional .05 level of significance.

The power of VNIS total scores to discriminate between a therapist's high and low outcome case is further illustrated by the data presented in Table 3.

Table 3

It may be seen that total VNIS scores, summed across raters for each therapist's high and low outcome cases correctly classify three therapist's high and low changers. The scores for the remaining two therapists are essentially the same. There were no instances of clear misclassification.

*t based on the separate variance estimate for the population variance for the difference in population means (see Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

Conclusions and Discussion

Taken together, these preliminary results suggest that the VNIS is a promising method for detecting signs of ineffective and harmful psychotherapy early in a therapeutic relationship. In addition, our findings have implications for research on the potential negative impact of specific therapeutic techniques. The low rater agreement on the Errors in Technique subsection confirmed earlier observations that what is viewed as an error in technique from one theoretical perspective may be judged competent or even good technique when viewed from some other theoretical position. This lack of consensus in the field regarding appropriate, effective techniques makes the construction of a pantheoretical instrument exceedingly problematic. The VNIS in its present form seems to apply most directly to time-limited dynamic therapy.

Regarding the issue of inter-rater reliability, the VNIS rating task involves an ordinal scale nested within a nominal scale. As Duncan and Fiske (1977) have pointed out, this increases the difficulty of obtaining adequate inter-judge consensus because two judgments are required rather than one. Furthermore, in training clinical raters, their theoretical backgrounds and biases, the type of clients they are most familiar with, and previous experience in making clinical ratings must be taken into account. The problem of inter-rater reliability, however, goes beyond the nature of the instrument and the selection and training of clinical raters. We believe that commonly accepted statistical methods fail to adequately reflect true rater consensus. More broadly, we see the need for systematic research on human observer variables that affect the precision and replicability of complex judgments of the kind demanded by the VNIS.

In the next phase of our research, we plan to rate therapy tapes from the files of other psychotherapy research projects. In view of the foregoing issues, we feel that the most productive next step is examination of the predictive power of the VNIS when it is used by raters who share both a theoretical framework and implicit standards for major rating decisions (e.g., what constitutes "normal limits").

In conclusion, our work has sharpened our awareness of the methodological problems that arise when clinical judgment is an integral part of a measurement technique. Although there may be definite upper limits to the precision that can be achieved with such an approach, we have seen considerable progress in our efforts to objectify factors widely-believed to lead to ineffective and harmful psychotherapy. In addition to endorsing the public sector's current demand for empirical evidence on the safety and efficacy of all psychosocial treatments (Parloff, 1980), research of the kind presented here may further our understanding of the process and outcome of psychotherapy.

References

Duncan, S. & Fiske, D. W. Face-to-face interaction: Research, methods, and theory. Hillsdale, New Jersey: Lawrence Erlbaum, 1977.

Finn, R. H. A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 1970, 30, 71-76.

Finn, R. H. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. Educational and Psychological Measurement, 1972, 35, 255-265.

Gomes-Schwartz, B. Development of a scale to predict negative change in psychotherapy. Paper presented at the Conference of the Society for Psychotherapy Research, Toronto, June, 1978.

Kiesler, D. J. An interpersonal communication analysis of relationship in psychotherapy. Psychiatry, 1979, 42, 299-311.

Nie, N. H., Hull, C. H., Jenkins, J. G., Stinbrenner, K., & Bent, D. H. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Parloff, M. B. Psychotherapy and research: An anaclitic depression. Psychiatry, 1980, 43, in press.

Strupp, H. H., Hadley, S. W., & Gomes-Schwartz, B. Psychotherapy for better or worse: The problem of negative effects. New York: Jason Aronson, 1977.

Tinsley, H. E. A. & Weiss, D. J. Interrater reliability and agreement of subjective judgments. Journal of Conseling Psychology, 1975, 22, 358-376.

TABLE 1

VANDERBILT NEGATIVE INDICATORS SCALE (VNIS)

SECTION	CATEGORY
I PATIENT PERSONAL QUALITIES & ATTITUDES (17 ITEMS)	LACK OF MOTIVATION
	INSUFFICIENT SELF-DISCLOSURE
	VERBAL COMMUNICATION PROBLEMS
	PROBLEMS WITH AFFECT
	SELF-DEROGATION
	UNREALISTIC EXPECTATIONS
	NEGATIVE ATTITUDES
	DEFICIENT EGO RESOURCES
II THERAPIST PERSONAL QUALITIES & ATTITUDES (9 ITEMS)	DEFICIENCIES IN THERAPEUTIC COMMITMENT
	REJECTING ATTITUDES
	CONTROLLING TENDENCIES
	DEFICIENT ENACTMENT OF THERAPIST'S ROLE
III ERRORS IN TECHNIQUE (10 ITEMS)	FAILURE TO MAKE NECESSARY INTERVENTIONS
	INAPPROPRIATE OR INADEQUATE INTERVENTIONS
	POTENTIALLY HARMFUL INTERVENTIONS
	INFLEXIBLE USE OF THERAPEUTIC TECHNIQUES
IV PATIENT-THERAPIST INTERACTION (2 ITEMS)	POOR THERAPEUTIC RELATIONSHIP
	POOR COLLABORATION
V GLOBAL SESSION RATINGS (4 ITEMS)	POOR MATCH BETWEEN PATIENT CHARACTERISTICS AND THERAPEUTIC APPROACH
	DULL INTERACTION
	NEGATIVE IMPACT OF THERAPY SESSION

TABLE 2

INTER-RATER RELIABILITIES[a]

SECTION	PEARSON R
PATIENT QUALITIES	.83
THERAPIST QUALITIES	.84
ERRORS IN TECHNIQUE	- .25
P-T INTERACTION	.48
TOTAL SCALE	.71
TOTAL SCALE EXCLUDING ERRORS IN TECHNIQUE	.82

NOTE. N = 10 FOR ALL CORRELATIONS.

[a]COMPUTATIONS BASED ON SUMS OF RELIABLY RATED ITEMS IN EACH SUBSECTION: PATIENT QUALITIES, N = 10, THERAPIST QUALITIES, N = 2; ERRORS IN TECHNIQUE, N = 5, P-T INTERACTION, N = 1. TOTAL SCALE N = 18.

TABLE 3

COMPARISON OF VNIS TOTAL SCORES[a] FOR EACH THERAPIST'S HIGH AND LOW OUTCOME CASES

	High Changer	Low Changer
Therapist 1	31	60
Therapist 2	48	127
Therapist 3	15	51
Therapist 4	34	29
Therapist 5	60	54

Note. Higher scores indicate more negative qualities.

[a]Figures are VNIS total scores on all 42 items, summed across raters.